对抗幻觉：垂直领域中大语言模型的应用策略探讨
–以中医知识问答领域为例

陈静[1]，曹智勋[1]

[1](华中师范大学信息管理学院 武汉 430079)

摘要：

[目的]本文旨在以中医知识问答领域为例，分析以知识库资源为例的非结构化知识和以知识图谱资源为例的结构化知识在提升大语言模型对抗幻觉效果上的差异性，并基于此进一步探讨大语言模型在垂直领域对抗幻觉能力的提升策略。

[方法]设计实验利用外部知识配合提示工程的方法，在中医知识问答领域进行知识库资源和知识图谱资源在提示效果上的差异性分析，并探讨动态三元组策略和融合微调策略等进行大语言模型对抗幻觉优化的优越性。

[结果]实验结果表明与知识库非结构化知识提示相比，知识图谱结构化知识提示在准确率、召回率和 F1 值方面表现更佳，分别比知识库提示高出 1.9%、2.42%和 2.2%，为 71.44%、60.76%和 65.31%；基于此进行了进一步优化策略分析后发现，动态三元组策略融合微调后在对抗幻觉上效果最佳，准确率、召回率和 F1 值分别达到了 72.47%、65.87%、68.62%。

[局限]本文的研究领域单一，目前只在中医问答领域进行了测试，尚需在广泛科研领域验证其泛化能力。

[结论]本研究证明了在中医领域，知识图谱结构化知识在减少幻觉现象和提升模型回复准确性方面优于传统非结构化知识，揭示了结构化知识在增强模型理解能力中的关键作用；微调策略和知识资源的融合使用为大语言模型提供了一种有效的性能提升路径。本文为大语言模型融合外部知识以提升知识服务提供了理论依据和方法支持。

关键词：对抗幻觉；大语言模型；提示工程；知识图谱；中医问答

分类号：G250 TP391

Combating Hallucinations: Application Strategies of Large Language Models in Vertical Domains - A Case Study in the Field of Traditional Chinese Medicine

Chen Jing[1], Cao Zhixun[1]

[1](School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract:

[Objective] This paper aims to analyze the differences in combating hallucinations in large language models between unstructured knowledge exemplified by knowledge base resources and structured knowledge exemplified by knowledge graph resources, using the Traditional Chinese Medicine (TCM) Q&A domain as a case study. It further discusses strategies to enhance the capability of large language models to combat hallucinations in vertical domains based on these findings.

[Methods] The study designed experiments using external knowledge combined with prompt engineering techniques to analyze the differences in prompting effects between knowledge base resources and knowledge graph resources in the TCM Q&A domain. It also explores the superiority of dynamic triplet strategies and integrated fine-tuning strategies in optimizing large language models against hallucinations.
[Results] Experimental results show that compared to prompts from unstructured knowledge in the knowledge base, prompts from structured knowledge in the knowledge graph perform better in terms of accuracy, recall, and F1 score, improving by 1.9%, 2.42%, and 2.2% respectively, reaching 71.44%, 60.76%, and 65.31%. Further analysis of optimization strategies revealed that the combination of dynamic triplet strategy and fine-tuning yielded the best effects against hallucinations, achieving accuracy, recall, and F1 scores of 72.47%, 65.87%, and 68.62%, respectively.
[Limitations] This study is limited to a single field, having been tested only in the domain of Traditional Chinese Medicine Q&A, and its generalizability needs to be validated in a broader range of scientific fields.
[Conclusions] This study has demonstrated that in the field of Traditional Chinese Medicine, structured knowledge from knowledge graphs surpasses traditional unstructured knowledge in reducing hallucinations and enhancing the accuracy of model responses. It reveals the critical role of structured knowledge in boosting model comprehension abilities; the integration of fine-tuning strategies with knowledge resources provides an effective pathway for performance enhancement in large language models. This paper provides theoretical justification and methodological support for integrating external knowledge into large language models to enhance knowledge services.
Keywords: Combating hallucinations; Large language models; Prompt engineering; Knowledge graphs; Traditional Chinese Medicine Q&A.

1 引言

人工智能的迅猛发展，为信息检索[1]、知识问答[2]和决策支持系统[3]带来了前所未有的变革。传统问答系统存在用户口语化表达严重[4]、回答质量参差不齐[5]等问题。随着各类大型语言模型（Large Language Model， LLM），如 GPT[6]、BERT[7]等的不断涌现，其丰富的预训练知识赋予了它们在通用领域展现强大能力的机会。然而大语言模型出现的“幻觉”问题，即模型生成的内容并非基于事实或准确信息[8]，使得用户难以信任其生成的回复，这种现象在一些对于专业知识有严格要求的垂直领域如医疗[9]、法律[10]等领域更为凸显。因此，对抗幻觉成为了大语言模型研究中的热点与难点。

针对垂直领域的幻觉问题，现有研究聚焦于利用外部知识来提升模型的准确性[11]。少量研究初步表明，不同形式的知识资源，如非结构化知识和结构化知识在对抗幻觉的效果可能存在差异。如 Ram 等的研究表明，如果外部知识是从结构化数据源、数据库或知识图谱中精心组织的，那么大语言模型的回复将更符合事实[12]。然而当前研究尚未深入探讨并实证这些知识资源在对抗大语言模型幻觉方面的效果差异，尤其对于需要高度专业知识和精确信息的领域，缺乏识别和选择对特定垂直领域最有价值的知识资源以及通过这些资源进一步优化模型性能的系统性研究。

大语言模型在特定专业领域，如医学领域的表现往往因缺乏领域专业知识而受到限制[13]。中医，作为一门拥有丰富理论体系和复杂概念的古老医学体系，在

全民健康中发挥着重要作用[14]，随着人们对综合医疗和中西医结合的兴趣增加，对中医知识的需求也日益增长。与西医相比，大语言模型在融入中医药领域方面尚未取得显著进展，主要因为中医领域公开资源的缺乏。这也导致大语言模型在处理中医相关问题时性能不佳[15]。因此，中医领域为测试和优化大语言模型在垂直专业领域应用的能力提供了理想的测试场景。

针对大语言模型对抗幻觉的问题，本文以外部知识库和三元组知识图谱为切入点，结合提示工程（Prompt Engineering）技术探索两类典型知识资源对提升模型效果的差异性。基于该差异结果，本文将进一步探索对抗幻觉的优化策略。具体而言，本文考虑了 Liu 等[16]提到的句子长度对模型回复精度的影响，通过动态策略调整三元组数量，将问句长度和三元组相似度结合起来优化三元组的选择，控制输入句子的长度。此外，根据已有研究表明[17]，微调可以显著提升模型在特定任务或领域的表现，本文结合外部知识和微调策略，进一步增强语言模型对抗幻觉的能力。本文的主要贡献有：

（1）以中医问答为例，探讨了知识库非结构化知识和知识图谱结构化知识对抗大语言模型幻觉的差异，优化了知识资源注入大语言模型的方法。进而，引入问句长度和三元组相似度得分两个变量来动态调整三元组数量，并探讨这种动态调整策略对大语言模型回复精度的影响，为理解问句长度和内容相关性对模型性能的影响提供了新的视角。

（2）在两种知识资源提示的基础上融入了微调策略，创新性地结合了微调技术和外部知识提示策略，以进一步提升大语言模型在回答特定领域问题时利用知识资源进行回复的准确性，精准地引入与问题相关的外部知识，有效地增强了模型对抗幻觉的能力。

2 研究现状

2.1 对抗幻觉策略相关研究

幻觉（hallucination），即大语言模型通常会产生一些看似合理但与认知无关或与事实不符的陈述[18]。“幻觉”的产生是因为大语言模型在生成输出文本时遵循了最大似然原则，而没有考虑事实性。因此，它们最终生成的输出总是高度可能，但不一定符合事实[19]。幻觉问题在医疗、法律等专业垂直领域表现得尤为突出，因为这些领域对信息的准确性和可靠性有着极高的要求。

现有研究利用不同的方法旨在降低大语言模型产生幻觉的可能性，Sun 等[20]利用对比学习调整参数优化大语言模型的隐性知识激发过程，从而减少他们在对话中的幻觉；Shi 等[21]通过参数调整和语境感知解码等技术手段降低了幻觉产生。但在面对大规模参数的模型时，这些方法的成本和灵活性受到限制，应用于垂直领域需要针对不同领域调整参数，进一步增加了成本。Wei 等[22]通过一种思维链（CoT）提示来激发模型推理能力，试图降低幻觉问题；Zhao 等[23]在 Wei 等的基础上提出了一个由 CoT 提示的验证与编辑框架，该框架根据外部检索到的知识对推理链进行后编辑，从而提高预测的保真度。尽管这些研究取得了一定的进展，但在严谨性要求较高的专业领域应用仍显示出局限性。Stiennon 等[24]通过模型预测人类偏好，并将其作为奖励函数，利用强化学习进行微调；Menick 等[25]通过从人类偏好中进行强化学习，选择少数问题拒绝回答，从而显著提高系统的可靠性。但是强化学习往往存在奖励函数不完善的问题。还有一种方法是利用“心智社会”，即协同多种语言模型的回复，Guerreiro 等[26]为了解决不同翻译场景下

大规模多语言模型中的幻觉问题，提出当原始系统出现幻觉时，可以请求其他翻译系统充当备用系统，但是多个语言模型协作会降低回复效率，且不同模型可能具有不同的知识和理解，这可能导致结果之间的一致性问题，无法应用于专业性强的垂直领域。

针对垂直领域，利用外部知识来对抗幻觉已逐渐受到研究者的关注。如，Borgeaud等[27]发现，将外部知识引入到大语言模型中可以有效提升模型的性能；Tam 等[28]也指出对于高度依赖精确信息的应用场景，外部知识的结合不仅能提升模型的准确率，还能显著减少生成的幻觉内容。

2.2 外部知识在垂直领域对抗幻觉的应用

在特定垂直领域内，模型需要理解并处理专业术语和复杂概念。外部知识库提供了这些领域特定语境的必要背景知识，使模型能更好地理解和生成与该领域相关的内容。通过整合大型文本数据库等外部知识，可以强化大语言模型在专业领域中的应用性能[29]。例如，Trautmann 等[30]将法律文档通过提示工程引入了大语言模型，以提高其在法律判决预测任务中的性能；Cui 等[31]引入了一种将矢量数据库检索与关键词检索相结合的方法来克服参考数据检索时法律数据筛选中的模型幻觉问题；张鹤译等[32]基于 Langchain 框架构建了中医方剂问答系统，通过自建知识库来生成更具备专业知识的回答；Peng 等[33]将从知识库中检索到的知识纳入提示中，使用文本嵌入进行检索，然后利用大语言模型的自动特征提取功能，以提高害虫识别任务的准确率。

上述研究多是利用知识库非结构化知识作为外部知识，而 Ye 等[34]的研究强调了知识图谱的结构化知识在构建以实体为中心的微调指令中的潜力，表明结合知识图谱可以提高模型在事实结果关联方面的性能。Bao 等[35]提出了 DISC-MedLLM，利用知识图谱构建高质量数据集，从而训练大语言模型在端到端对话式医疗保健服务中提供准确、真实的医疗回复；Gui 等[36]通过知识图谱创建指令数据集来增强大语言模型在信息提取任务中的性能；Wang 等[37]以中医药知识图谱为基础，围绕核心实体及各种意图进行指令生成，以优化大语言模型在中医领域的能力。

还有学者将知识图谱和提示工程相结合：Baek 等[38]利用检索的方法检索知识图谱中与问题相关的三元组提示给大语言模型，以提高其对抗幻觉的能力；Wu 等[39]将知识图谱中提取的三元组重写为文本化的语句进一步提高大型语言模型的性能；Wen 等[40]结合思维导图赋予大语言模型最新知识并引导其生成推理路径，以减轻幻觉现象。

综上，在专业垂直领域内有效整合并利用外部知识的重要性已受到研究者的重视。然而，现有研究虽已探讨了不同形式的外部知识如非结构化知识和结构化知识在垂直领域中增强大语言模型性能的方法，但鲜有对不同知识资源在对抗大语言模型幻觉问题上的效果进行比较以分析其优越性，更少见基于外部资源的进一步优化策略研究。本文首先剖析了知识库和知识图谱两类典型的外部知识资源在提升语言模型对抗幻觉效果上的差异性；然后基于分析结果进一步探究大语言模型的优化方式。现有研究使用的优化策略包括架构创新、训练策略优化、上下文长度优化、模型微调等[41]，其中，上下文长度的改进和模型微调是这些策略诸多研究中较为流行且有效的两种策略[42]。这两种策略通过调整输入数据的长度和对模型进行特定领域的调整，显著提升了模型的性能。根据 Liu 等[16]的研究，句子长度对模型的回复精度有显著影响，因此，通过结合问句长度和三元组相似

度，本文采用动态策略优化三元组的选择。另一方面，结合微调策略在专业领域的有效应用[43]，本文利用外部知识结合微调策略进一步增强了大语言模型在特定任务或领域中的表现，特别是在对抗幻觉方面的能力，以提高大语言模型在垂直领域应用的性能和实用性。

3 研究设计

3.1 研究框架

本文首先将基于知识库和知识图谱两类典型的外部知识资源，运用提示工程探究二者对提升大语言模型在垂直领域对抗幻觉效果的差异性。研究思路主要分为如下三个部分：知识提取、比较分析、提升策略。

在知识提取阶段：对于知识图谱资源，本研究使用命名实体识别获取问题中的实体，将识别出的实体在节点词典中进行匹配，最后在知识图谱中查询相关三元组，得到结构化知识；对于知识库资源，将输入的问题转化为嵌入向量，并在知识库文件构成的向量数据库中进行搜索，以获取非结构化知识。将上述得到的结构化和非结构化知识利用提示模板，为输入问题构建精确的提示词。

在比较分析阶段：使用中医知识问答数据集，探究不同知识资源在提高模型对抗幻觉能力的效果差异。针对中医问题，通过知识提取和模板构建方法形成不同知识资源构建的提示语句，然后将提示语句和问题一同输入至大语言模型，从而生成相应的回答。最后利用 BERTScore 指标[53]中的准确率、召回率、F1 得分进行对比评估，以选择最优知识资源。

在提升策略阶段：根据 Liu 等的研究[16]，对于效果更优的结构化知识，采用一种动态策略，通过输入问句的长度以及三元组与问句之间的相似度来进一步优化三元组数量，提高数据的相关性；结合微调对专业领域大语言模型能力的提升，对于大语言模型，采用了专门的微调数据集对其进行微调，从而提升模型对外部知识的理解，增强其对抗幻觉的能力。具体研究流程如图 1 所示。

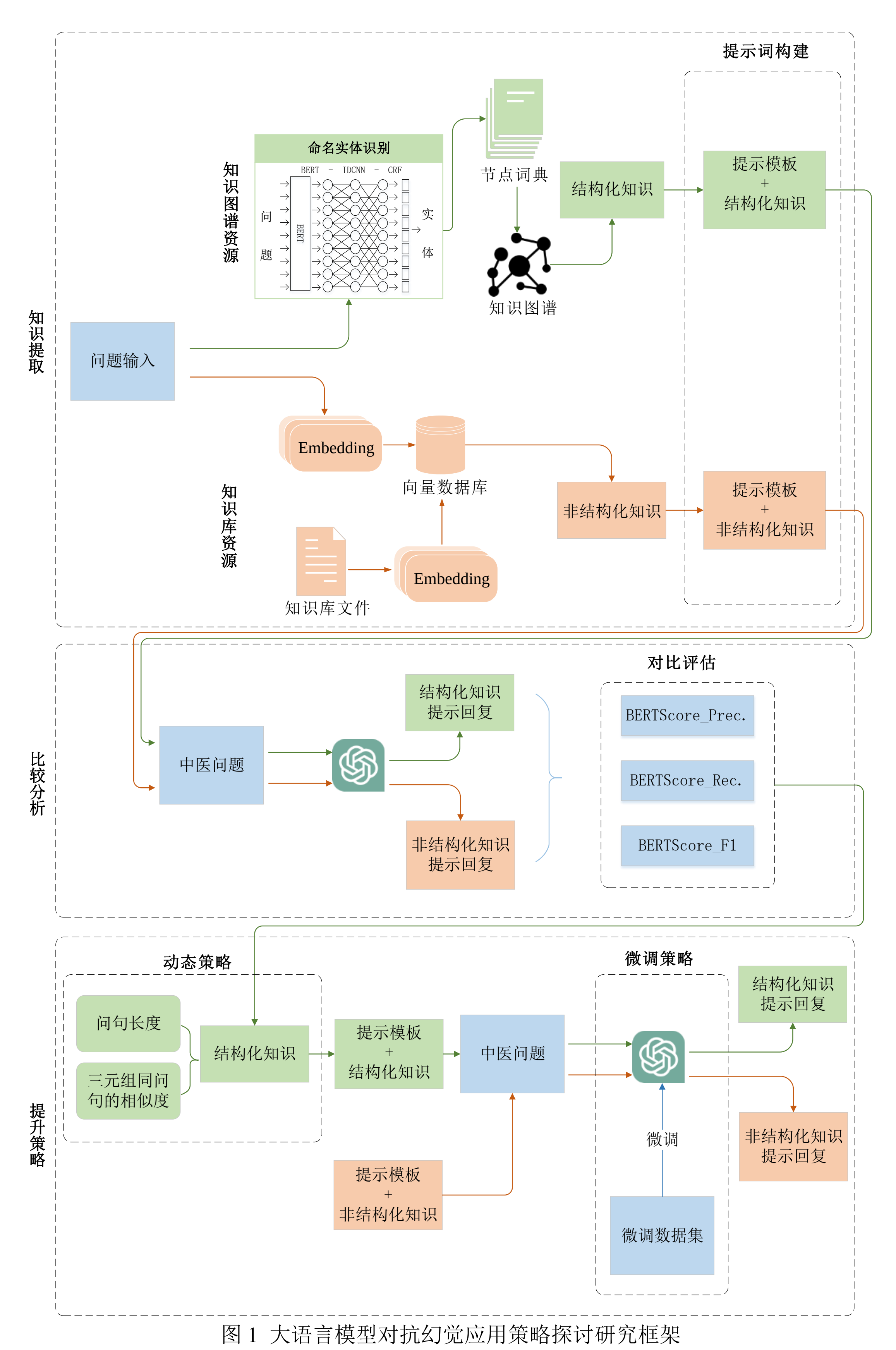


图 1 大语言模型对抗幻觉应用策略探讨研究框架

Fig.1 Exploring the Framework for Application Strategies Against Hallucinations in Large Language Models

3.2 数据收集与预处理

本文所使用的问答数据集为阿里天池中医文献问题生成数据集[1]，数据源来自中医药领域文本，包括【黄帝内经翻译版】、【名医百科中医篇】、【中成药用药卷】、【慢性病养生保健科普知识】四个主要来源，共标注 13000 对（问题、文档、答案），来源于 5000 篇文档，每篇文档由人工标注产生 1～4 对(问题，答案)对。仅保留数据集中的问题和答案，得到 18478 条问答对。由于本文意在证明策略的有效性，因此需要删去不满足要求的数据，并去除命名实体识别中未识别出实体、节点查询中没有查询到节点以及查询得到的三元组数少于 10 个的问答对，最终保留 4000 条数据。考虑到 token 和算力的限制，本文采取类似白如江等[44]的数据选择方式进行小规模验证，随机选取其中的 500 条作为本文数据。

3.3 外部知识增强大语言模型提示方法

（1）知识提取

①知识图谱增强的大语言模型提示

本小节主要介绍利用知识图谱增强大语言模型的细节。知识图谱$\mathcal{G}$由一组事实三元组$t = \{(s, r, o)\}$构成，其中s和o表示主观实体和客观实体，r表示二者之间的关系[38]。

利用知识图谱构建提示的核心在于如何从知识图谱中获取与问题相关、对回答问题有帮助的三元组。本文通过命名实体识别技术识别问题中的实体，并结合 Aho-Corasick 算法和 Sentence-BERT 模型将实体与知识图谱节点匹配，利用 neo4j 图数据库的 cypher 查询语句提取与问题相关的三元组，最后将得到的三元组与问题进行相似度比较，按相似度得分对三元组进行排序。具体工作流程为：

1）命名实体识别（NER）。NER 是自然语言处理任务中最重要的任务之一，目前 NER 技术的主流方法是使用条件随机场（CRF），研究发现该方法可以利用句子级标签信息[45]。随着预训练模型如 BERT 等[7]的发展，有学者将 BERT 融入 NER 中并取得了更好的结果[46]，本文使用 BERT-IDCNN-CRF 模型[47]，利用 BERT 进行句子级特征提取，应用于中医问答句子的实体识别，并得到实体集$E = \{e_1, e_2, \dots, e_n\}$。

使用阿里天池中医药说明书数据集[2]作为训练数据集，并将其转化为 BIEO 格式对 BERT-IDCNN-CRF 模型进行训练，以 2e-05 的学习率和 24 的批量大小训练了 3 个轮次，最终模型在验证集上取得了 86%的 F1 得分，使用该模型进行中医知识问答中的实体识别。

2）节点匹配。由于构建新的知识图谱十分复杂，本文使用现有的中医知识图谱 TCMKG[3],该知识图谱中包含 37000 个实体和 120000 个关系，经过筛选后最终保留 19000 个实体和 55000 个关系。将知识图谱中的节点n_i建立为一个词表$\mathcal{D}$，利用 Aho-Corasick 算法将 NER 识别出的实体$e_i \in E$与节点$n_i \in \mathcal{D}$相匹配，初步得到匹配节点集N。为了避免同义词未匹配的情况，如“头痛”和“头疼”，本文还使用 Sentence-BERT 计算实体和节点的相似度得分，并将得分在 0.9 以上的实体

---

[1] https://tianchi.aliyun.com/dataset/86895

[2] https://tianchi.aliyun.com/dataset/86819

[3] https://github.com/ywjawmw/TCM_KG

进行保留，以确保能得到所有与问题相关的三元组。即对于每个$e_i \in E$和$n_i \in N$使用 Sentence-BERT 计算相似度得分$S_{e-n}(e,n)$，若$S_{e-n}(e,n) > 0.9$，则n被认为是与e高度相似的节点，并保留在最终的节点集合N^*中，公式表示为：

$$N^* = \{ n \in N \mid S_{e-n}(e,n) > 0.9, \forall e \in E \} \quad (1)$$

其中E为 NER 识别出的实体集合，N是通过 Aho-Corasick 算法匹配到的初步节点集合，$S_{e-n}(e,n)$是使用 Sentence-BERT 计算得到的实体e与节点n之间的相似度得分，N^*为最终保留的知识图谱节点集合。

3）三元组查询。将实体与节点匹配后保留下来的节点集N^*利用 neo4j 图数据库的 cypher 查询语句进行查询，返回结果为被查询节点、该节点的邻居节点、两个节点之间的关系，共同形成的三元组t，对于每个问题q都有一个与其对应的三元组集$T = \{t_1, t_2, \dots, t_n\}$。

4）相似度排序。为了进一步筛选与问题相关的三元组，本文参考文档检索的方法，即根据文档的嵌入相似性检索与给定查询相关的文档。将原始问句$Q = \{q_1, q_2, \dots, q_n\}$和经过查询后得到的三元组总集合$\mathbb{T} = \{T_1, T_2, \dots, T_n\}$利用现有句子嵌入模型$M$嵌入到向量空间，之后计算二者的相似度得分。公式如下：

$$V_{q_i} = M(q_i),\ V_{t_i} = M(t_i), \forall t_i \in T_i \quad (2)$$

$$S_{q-t}\left(V_{q_i}, V_{t_i}\right) = \cos\left(V_{q_i}, V_{t_i}\right), \forall t_i \in T_i \quad (3)$$

$$T_{sorted} = sort\left(T, key = S_{q-t}\left(V_{q_i}, V_{t_i}\right)\right) \quad (4)$$

其中，q_i是原始问题句子，T_i是q_i对应的三元组集，t_i是T_i中的三元组，M是句子嵌入模型，V_{q_i}是问题句子向量表示，V_{t_i}是第i个三元组的向量表示，$S_{q-t}\left(V_{q_i}, V_{t_i}\right)$是$V_{q_i}$和$V_{t_i}$的相似度得分，$T_{sorted}$是根据相似度得分进行降序后的三元组集合。

使用的嵌入模型为 huggingface 提供的 distiluse-base-multilingual-cased-v1 模型[4]。最后根据三元组与原始问题的相似度得分进行降序排序。整体流程如图 2 所示：

[4] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

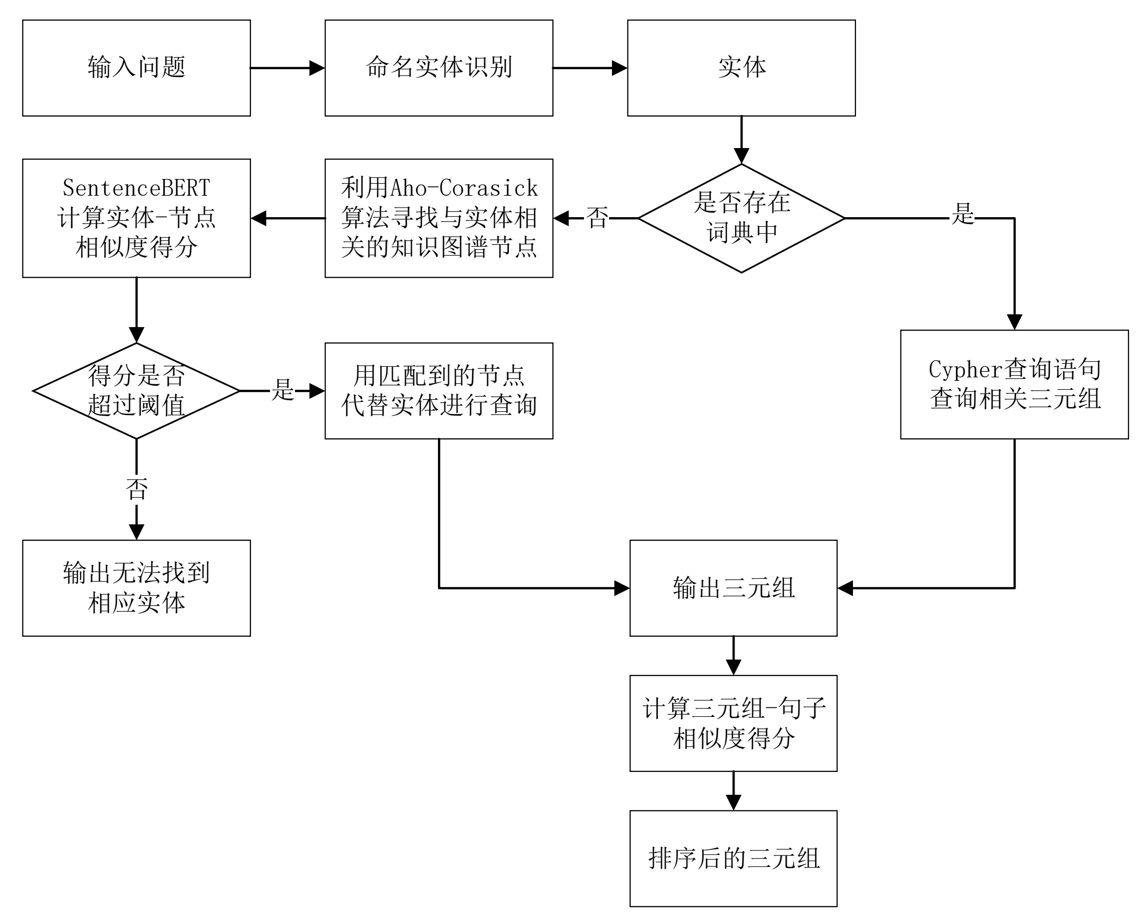


图 2 知识图谱三元组获取流程图

Fig. 2 Knowledge Graph Triple Extraction Process Flowchart

②知识库增强的大语言模型提示

现有外部知识主要包括知识库非结构化知识和知识图谱结构化知识。本小节主要介绍如何利用知识库非结构化知识增强大语言模型。Langchain[5]是一个开源框架，允许开发者将大语言模型和数据来源结合起来。张鹤译等[32]利用 Langchain 框架建立起知识库和大语言模型间的链接，参考其研究，本文主要使用 Langchain 的文档检索框架，对知识库进行检索，以获取与问题相关的文本知识资源。

本文使用古书、古籍等共计 700 项中医药古籍文本[6]作为知识库来源，并利用 Langchain 框架将文件进行处理，嵌入向量空间形成向量数据库。在该框架下，对知识库中的每个文件$F_i(i = 1,2,3,\dots,n)$进行处理，将其分解为若干个文本块$D_{ij}(i = 1,2,3,\dots,n; j = 1,2,3,\dots,m)$其中$i$为文件编号，$j$为文本块编号。接下来为每个文本块创建向量索引$V_i(i = 1,2,3,\dots,n \times m)$，并存放在向量数据库中，将原始问句$q_i$转化为向量$V_{q_i}$，通过计算$V_{q_i}$与每个$V_i$的相似度，找到最相似的$k$个文本块的向量索引，并拼接起来形成最终的文本$w$，对于每个$q_i$都有一个文本$w_i$对应。

[5] https://www.langchain.com/

[6] https://github.com/xiaopangxia/TCM-Ancient-Books

③知识注入

为了保证知识库文本知识和知识图谱结构化知识输入的一致性，本文将二者均进行线性语言化并通过构建模板$TEMP$将外部知识注入大语言模型，使其能以输入的外部知识为基础输出答案，如下表 1 所示。关于模板$TEMP$的构建具体为：对于知识图谱结构化知识，针对问题q有对应的三元组集$T=\{t_1,t_2,\dots,t_n\}$，附加说明："下面是三元组形式事实："，使用$TEMP_{kg}:T\rightarrow T'$，最终数学化表示为$P(y|[T',q'])$。类似的，对于知识库文本知识，针对问题$q$有对应的文本$w$，构建过程中附加一条指示性说明："下面是事实资料："则$TEMP_{kb}:w\rightarrow w'$，这样形成的$w'$会被并入$q'$，最终大语言模型生成答案的过程可以数学化表示为$P(y|[w',q'])$。

表 1 知识注入案例

Table 1 Knowledge Injection Case Study

问题(q)	外部知识	提示模板($TEMP$)	Prompt
决明子有什么功效？	$\boldsymbol{T}=$"（决明子，功能，清热明目）、（决明子，功能，润肠通便）……"	$\boldsymbol{TEMP_{kg}}=$"下面是三元组形式事实："	$\boldsymbol{T'}=$"下面是三元组形式事实：（决明子，功能，清热明目）、（决明子，功能，润肠通便）……"
	$\boldsymbol{w}=$"属性：主治泻肝邪治头风。及一切目疾。尚有草决明石决明二种。功用相同。附载于后。 ……"	$\boldsymbol{TEMP_{kb}}=$"下面是事实资料："	$\boldsymbol{w'}=$"下面是事实资料：属性：主治泻肝邪治头风。及一切目疾。尚有草决明石决明二种。功用相同。附载于后。 ……"

（2）比较分析

本文利用中医知识问答数据集来分析不同知识资源在增强模型抗幻觉能力方面的表现差异。通过知识提取模块生成了基于不同知识资源的提示语句，包括基于知识图谱的$\boldsymbol{T'}$以及基于知识库的$\boldsymbol{w'}$。这些提示语句与具体问题一起输入到大型语言模型生成模型回复A'。通过将模型生成的回答 A'与原始答案A进行比较，并计算各项性能指标得分，本研究对比评估了不同知识资源的效果，以确定最为有效的知识资源。

（3）提升策略

经过比较分析后，为进一步提大语言模型对抗幻觉的能力，本文提出了包括动态策略和外部知识融合微调的策略。动态策略通过实时分析输入问句的长度和复杂度，调整三元组选择的标准，确保选取的三元组与问句高度相关，从而减少模型生成不相关或错误信息的可能性；外部知识融合微调的策略，通过对模型进行微调，增强模型对外部知识的理解能力。

① 动态三元组策略

为深入探究三元组数量对模型回复的影响，本文根据问句长度$len(q)$以及三元组和问句的相似度得分S_{q-t}，进一步对三元组数量进行控制。具体来说，通过两个变量控制三元组数量，分别是根据问句长度确定的最大输入数量Max_K和根据相似度得分确定的输入三元组百分比$K_percent$，二者定义公式为：

$$Max_{K_{qi}} = \frac{len(q_i) - \min(Q) * 5}{\max(Q) - \min(Q)} + 5 \tag{5}$$

$$K_{percent_{qi}} = median\left(S_{q-t}\right) * 100\% \tag{6}$$

其中，$len(q_i)$为第i个问句的长度，$\min(Q)$为所有问句长度的最小值，$\max(Q)$为所有问句长度的最大值，$median\left(S_{q-t}\right)$为所有三元组得分的中位数。

②外部知识融合微调策略

为进一步探究大语言模型效果改善策略，本文将模型微调与知识资源提示相结合。具体而言，利用 LoRa 微调[48]的方式对 ChatGLM3 模型进行微调，使用的训练数据集为上文提到的 18478 条数据并去除命名实体识别筛选的 4000 条数据，分为训练集 11582 条数据，测试集 2896 条数据，训练超参数设置如表 2 所示：

表 2 LoRa 微调实验超参数设置

Table 2 LoRa Fine-tuning Experiment Hyperparameter Settings

参数	参数值
Epoch	4
Batch_size	4
LoRa Rank	8
Learning Rate	5e-5

将训练得到的 LoRa 权重参数替换原 ChatGLM3 模型的参数。

3.4 语言模型的选择

为了补偿样本数量的限制和验证外部知识提示的有效性，本文采用各类不同参数量的大语言模型来确保实验结果的广泛适用性和鲁棒性。包括：

通义千问(1.8B,7B,14B,72B)[49]：是由阿里云构建的大型语言模型。针对多达 3 万亿个多语言数据进行了稳定的预训练，覆盖了广泛的领域、语言（以中文和英语为重点）等。

百川(7B,13B)[50]：由百川智能开发的一个开源可商用的大规模预训练语言模型。基于 Transformer 结构，在大约 1.2 万亿 tokens 上训练的 70 亿参数模型，支持中英双语。在标准的中文和英文 benchmark 上均取得同尺寸最好的效果。

ChatGLM3(6B)[51]：ChatGLM3 是智谱 AI 和清华大学 KEG 实验室联合发布的对话预训练模型。在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。

GPT3.5(175B)[52]：GPT3.5 是由 OpenAI 公司开发的语言模型，其在人工标注训练数据的基础上，再使用强化学习来增强预训练模型的能力。与之前的版本

相比，GPT-3.5 在规模上更大、参数更多，具备了更强的语义理解和文本生成的能力。

3.5 评估指标

对于大语言模型生成内容的评估，考虑到生成内容与原答案存在含义相同但表达差异的问题，不直接使用基于字符匹配的评估方法。本文使用 BERTScore[53]作为评估指标，BERTScore 是一种用于文本生成的自动评估指标，可计算候选句子中每个标记与参考句子中每个标记的相似度得分。它利用 BERT 模型预先训练的上下文嵌入，通过余弦相似度来匹配候选句和参考句中的单词，考虑了上下文的语义、语法信息，更具有鲁棒性。

和传统字符匹配指标类似，BERTScore 中同样包括 P（precision）、R（recall）、F1 值三个指标，各指标计算方法如公式所示：

$$P_{BERT}=\frac{1}{|\hat{x}|}\sum_{\hat{x}_j\in\hat{x}}\max_{x_i\in x}\mathrm{X}_i^T\hat{\mathrm{X}}_j \tag{7}$$

$$R_{BERT}=\frac{1}{|x|}\sum_{x_i\in x}\max_{\hat{x}_j\in\hat{x}}\mathrm{X}_i^T\hat{\mathrm{X}}_j \tag{8}$$

$$F_{BERT}=2\frac{P_{BERT}\cdot R_{BERT}}{P_{BERT}+R_{BERT}} \tag{9}$$

其中x为原始句子，x_i为组成该句子的 token，X为其嵌入向量；$\hat{x}$为比较的句子，$\hat{x}_j$为组成该句子的 token，$\hat{\mathrm{X}}$为其嵌入向量。

完整得分的计算方式将中x的每个 token 与$\hat{x}$中的一个 token 进行匹配，以计算召回率；将$\hat{x}$中的每个 token 与x中的一个 token 进行匹配，以计算精确度。使用贪婪匹配来最大化匹配相似度得分，其中每个 token 都与另一个句子中最相似的 token 进行匹配。将精确度和召回率结合起来计算 F1 指标。

4 实证研究

4.1 实验环境与参数设置

本实验环境配置为：CPU，12th Gen Intel(R) Core(TM) i7-12700H；GPU，NVIDIA GeForce RTX 3060 Laptop GPU；显存：6G；Python 版本，3.10.11；Cuda 版本，12.4。

对于注入的知识图谱三元组的数量，经过 Baek 等人[38]的实验，发现三元组数在 5-10 个之间时通用模型达到最佳性能，因此采用该团队的设置，将数量K设定为 10。

4.2 不同知识资源对大语言模型对抗幻觉效果的影响

为探究不同知识资源提示下大语言模型回复的效果差异，本小节通过比较无提示模型回复、知识库提示回复、随机选取的与问题相关的K个三元组以及固定选取相似度得分排名前K个三元组进行效果比较，结果如表 3 所示。其中，使用知识图谱的两组提示：随机三元组和固定选取的三元组在准确率、召回率、F1 值上的得分都要高于无提示回复和知识库提示回复，且固定选取的三元组得分最高，分别达到了 71.44%、60.76%、65.31%，相较于无提示回复分别上升了 1.52%、

4.44%、3.19%，这说明知识图谱的三元组结构化知识作为凝练后的知识，更易于大语言模型理解分析，使得模型生成的回复更加符合事实，进而提升了模型对抗幻觉的能力。而知识库非结构化知识提示的改进幅度较小，仅在召回率和 F1 值上有所提升，分别提升了 2.02%、0.99%，甚至在准确率得分上有所下降。可能的原因是知识库检索到的文本知识较长，其中的无关信息分散了大语言模型的注意力，导致模型性能的下降。总体来说，两种知识资源在平均效果上对大语言模型均有改善作用，相较于知识库非结构化知识，知识图谱结构化知识的效果更佳。

表 3 不同知识资源对语言模型影响的整体效果对比

Table 3 Comparative Overall Effect of Different Knowledge Resources on Language Models

评价指标 (BERTScore)	方法	qwen (1.8B)	qwen (7B)	qwen (14B)	qwen (72B)	baichuan (7B)	baichuan (13B)	gpt3_5 (135B)	chatglm3 (6B)	Average
P(%)	no prompt	69.82	70.40	69.80	**71.17**	69.53	70.22	69.39	69.07	69.92
	knowledge base	69.98	68.97	70.30	70.94	69.82	70.18	66.05	70.04	69.54
	random triple	71.87	71.04	71.58	69.84	69.99	70.93	71.53	71.26	71.00
	fixed triple	**72.24**	**71.05**	**72.13**	70.70	**70.50**	**71.20**	**71.96**	**71.73**	**71.44**
R(%)	no prompt	57.26	60.48	57.55	56.05	53.94	55.62	56.22	53.44	56.32
	knowledge base	62.64	**62.33**	59.50	57.17	55.47	56.28	57.01	56.33	58.34
	random triple	62.24	61.13	**62.67**	58.58	55.96	61.18	61.46	58.66	60.24
	fixed triple	**63.06**	60.13	62.32	**60.03**	**56.60**	**62.98**	**62.12**	**58.84**	**60.76**
F1(%)	no prompt	62.63	64.79	62.78	62.44	60.52	61.85	61.92	60.06	62.12
	knowledge base	65.71	65.12	64.04	63.02	61.58	62.22	60.97	62.19	63.11
	random triple	66.34	**65.37**	66.39	63.34	61.95	65.36	65.85	64.06	64.83
	fixed triple	**66.93**	64.83	**66.43**	**64.54**	**62.50**	**66.45**	**66.38**	**64.38**	**65.31**

注：粗体标注为最好结果，下划线标注为次好结果

4.3 动态策略与微调技术对模型性能的提升

（1）动态三元组策略对模型性能的影响

根据问句长度和相似度得分动态变化的三元组数与固定选取的K个三元组的效果对比，发现动态策略相较于固定K策略的提升效果很小，仅在召回率和 F1 值上提升了 0.64%、0.19%，可能的原因有：一、句子长度和相似度得分并非决定性因素：虽然句子长度和相似度得分在一定程度上反映了问题的复杂性和相关三元组的质量，但它们可能并不是影响模型回复精度的决定性因素；二、过拟合的风险：在某些情况下，过多依赖于与问题高度相似的三元组可能会使模型在特定问题上过拟合，而忽略了更广泛的语境理解和知识应用能力。具体原因还需在后续研究中探讨。

（2）模型微调与外部知识理解能力的提升

对 ChatGLM3 模型进行微调后，对比无提示回复、知识库知识、随机三元组、固定三元组、动态三元组 5 种策略微调前后的大语言模型回复效果，如图 3 所示，

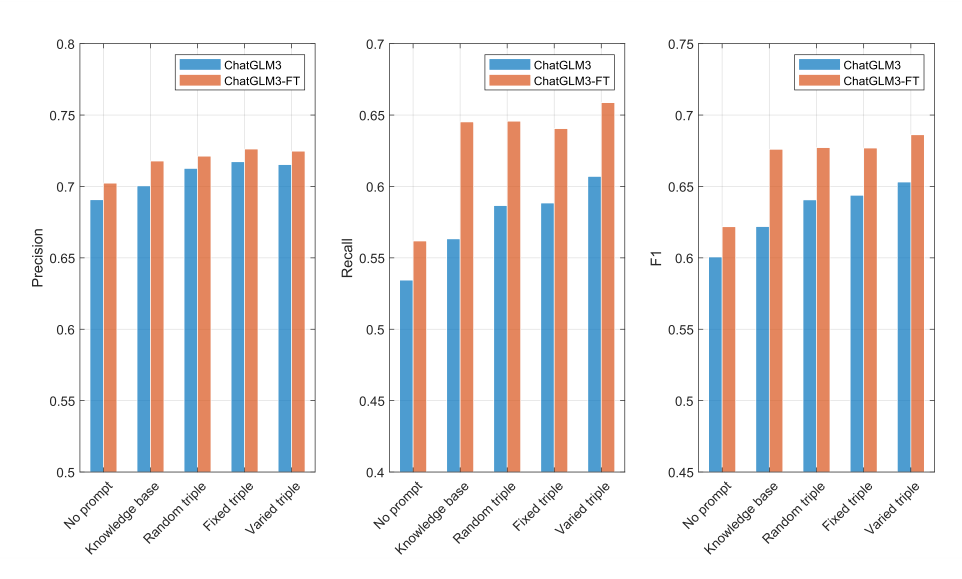


图 3 微调策略前后效果对比

Fig. 3 Comparison of Effects Before and After Fine-tuning Strategy

相较于原始模型，微调后的模型在各个策略的效果都有所提升，在 F1 值上各策略相较微调前提升了 2.12%、5.41%、3.66%、3.31%、3.31%，从提升幅度可以看出微调策略可以进一步帮助模型提高对两种外部知识的理解，且微调加随机三元组、固定三元组、动态三元组的三种策略的效果在 F1 值上达到了 67.72%，67.69%，68.62%要优于微调原始模型和微调加文本知识策略的 62.18%、67.60%，说明微调后的模型能更好地利用知识图谱的结构化知识。

在多种策略中，微调结合动态三元组的策略效果最佳，具体效果如下表 4 所示。

表 4 不同策略的指标得分对比

Table 4 Comparison of Metric Scores for Different Strategies

方法	P(%)	R(%)	F1(%)
ChatGLM3	69.07	53.44	60.06
ChatGLM3+FT	70.23	56.18	62.18
ChatGLM3+VT	71.53	60.70	65.31
ChatGLM3+FT+VT	**72.47**	**65.87**	**68.62**

注：粗体标注为最好结果　FT : FineTuning; VT : VariedTripe

ChatGLM3-FT+Varied_triple 在 F1 值上达到了 68.62%相较于 ChatGLM3、ChatGLM3-FT、ChatGLM3+Varied_triple 策略在 F1 得分上分别提升了 8.56%、6.44%、3.31%。这一结果表明微调后的模型在理解动态三元组策略提供的结构化信息方面表现更好。微调过程可能增强了模型对结构化格式知识的敏感性和适应性，从而提高了其学习能力。

5 结论

本研究旨在探索两类典型知识资源在提升大语言模型对抗幻觉效果的差异性，进一步探讨大语言模型在垂直领域对抗幻觉能力的提升策略。利用知识提取

模块为大语言模型引入知识库非结构化知识和知识图谱结构化知识两种知识资源；以中医知识问答领域为例，通过比较分析模块探索二者在提升大语言模型对抗幻觉效果上的差异性；最后基于差异结果利用动态策略和微调策略进一步优化大语言模型对知识资源的使用，进一步提升大语言模型对抗幻觉的能力。本研究不仅在技术层面开拓了大语言模型与外部知识融合的新途径，也在应用层面对模型对抗幻觉的能力进行了深入的探究。通过实验，证实了在中医领域知识图谱的结构化知识在减少幻觉现象、增强模型回应的准确度方面要优于传统的非结构化知识。这一发现对于大语言模型的理论构建和实践应用均具有重要意义。

在理论层面，本研究推动了对大语言模型认知机制的深入理解。通过揭示结构化知识在提升模型理解能力中的作用，为未来语言模型的外部知识融合提供了宝贵的经验。此外，本研究提出的微调策略和知识资源的融合使用为大语言模型在特定领域内提供了一种有效的性能提升路径。

在实践层面，本研究对中医问答系统等垂直领域应用提供了切实的技术支持。通过将信息密集的结构化知识应用于实际问题，模型能够给出更为准确的回复，推动了大语言模型在专业领域中的应用。

本文的研究内容还局限于一个领域和一种知识注入方法，未来的工作可以跨领域验证本研究的方法论，探索结构化知识与大语言模型融合对不同领域任务性能的影响。此外，还可以对知识注入的方法进行优化，提升大语言模型对抗幻觉的能力。

参考文献

[1] Wang J, Huang J X, Tu X, et al. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges[J]. ACM Computing Surveys, 2024, 56(7): 1-33.

[2] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.

[3] Solares E, De-León-Gómez V, Salas F G, et al. A comprehensive decision support system for stock investment decisions[J]. Expert Systems with Applications, 2022, 210: 118485.

[4] 王寅秋,虞为,陈俊鹏.融合知识图谱的中文医疗问答社区自动问答研究[J].数据分析与知识发现,2023,7(03):97-109.(Wang Yinqiu, Yu Wei, Chen Junpeng, et al. Automatic Question-Answering in Chinese Medical Q & A Community with Knowledge Graph[J]. Data Analysis and Knowledge Discovery,2023,7(03):97-109.)

[5] 易明,张婷婷.大众性问答社区答案质量排序方法研究[J].数据分析与知识发现,2019,3(06):12-20.(Yi Ming, Zhang Tingting, et al. Ranking Answer Quality of Popular Q&A Community[J]. Data Analysis and Knowledge Discovery,2019,3(06):12-20.)

[6] Radford A, Narasimhan K, Salimans T, et al.Improving Language Understanding by Generative Pre-training[J]. 2018.

[7] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J].arXiv preprint arXiv:1810.04805, 2018.

[8] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[9] Puchert P, Poonam P, van Onzenoodt C, et al. LLMMaps--A Visual Metaphor for Stratified Evaluation of Large Language Models[J]. arXiv preprint arXiv:2304.00457, 2023.

[10] Nay J J, Karamardian D, Lawsky S B, et al. Large language models as tax attorneys: a case study in legal capabilities emergence[J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270): 20230159.

[11] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens[C]//International conference on machine learning. PMLR, 2022: 2206-2240.
[12] Ram O, Levine Y, Dalmedigos I, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.
[13] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.
[14] Wang W Y, Zhou H, Wang Y F, et al. Current policies and measures on the development of traditional Chinese medicine in China[J]. Pharmacological research, 2021, 163: 105187.
[15] Yang G, Shi J, Wang Z, et al. TCM-GPT: Efficient Pre-training of Large Language Models for Domain Adaptation in Traditional Chinese Medicine[J]. arXiv preprint arXiv:2311.01786, 2023.
[16] Liu X, Ji K, Fu Y, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.
[17] Key O, Kaddour J, Minervini P. Local LoRA: Memory-Efficient Fine-Tuning of Large Language Models[C]//Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ NeurIPS 2023). 2023.
[18] Dash D, Thapa R, Banda J M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery[J]. arXiv preprint arXiv:2304.13714, 2023.
[19] Sadat M, Zhou Z, Lange L, et al. Delucionqa: Detecting hallucinations in domain-specific question answering[J]. arXiv preprint arXiv:2312.05200, 2023.
[20] Sun W, Shi Z, Gao S, et al. Contrastive learning reduces hallucination in conversations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13618-13626.
[21] Shi W, Han X, Lewis M, et al. Trusting your evidence: Hallucinate less with context-aware decoding[J]. arXiv preprint arXiv:2305.14739, 2023.
[22] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.
[23] Zhao R, Li X, Joty S, et al. Verify-and-edit: A knowledge-enhanced chain-of-thought framework[J]. arXiv preprint arXiv:2305.03268, 2023.
[24] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020, 33: 3008-3021.
[25] Menick J, Trebacz M, Mikulik V, et al. Teaching language models to support answers with verified quotes[J]. arXiv preprint arXiv:2203.11147, 2022.
[26] Guerreiro N M, Alves D M, Waldendorf J, et al. Hallucinations in large multilingual translation models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1500-1517.
[27] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens[C]//International conference on machine learning. PMLR, 2022: 2206-2240.
[28] Tam D, Mascarenhas A, Zhang S, et al. Evaluating the factual consistency of large language models through news summarization[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 5220-5255.
[29] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens[C]//International conference on machine learning. PMLR, 2022: 2206-2240.
[30] Trautmann D, Petrova A, Schilder F. Legal prompt engineering for multilingual legal judgement prediction[J]. arXiv preprint arXiv:2212.02199, 2022.
[31] Cui J, Li Z, Yan Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092, 2023.
[32] 张鹤译,王鑫,韩立帆等.大语言模型融合知识图谱的问答系统研究[J].计算机科学与探

索,2023,17(10):2377-2388. (Zhang Heyi, Wang Xin, Han Lifan, et al. Research on Question Answering System on Joint of Knowledge Graph and Large Language Models[J]. Journal of Frontiers of Computer Science & Technology, 2023, 17(10): 2377-2388.)

[33] Peng R, Liu K, Yang P, et al. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data[J]. arXiv preprint arXiv:2308.03107, 2023.

[34] Ye H, Liu T, Zhang A, et al. Cognitive mirage: A review of hallucinations in large language models[J]. arXiv preprint arXiv:2309.06794, 2023.

[35] Bao Z, Chen W, Xiao S, et al. Disc-medllm: Bridging general large language models and real-world medical consultation[J]. arXiv preprint arXiv:2308.14346, 2023.

[36] Gui H, Zhang J, Ye H, et al. Instructie: A chinese instruction-based information extraction dataset[J]. arXiv preprint arXiv:2305.11527, 2023.

[37] Wenjing Yue Wei Zhu and Xiaoling Wang. 2023. Shennong-tcm: A traditional chinese medicine large language model. https://github.com/ michael-wzhu/ShenNong-TCM-LLM

[38] Baek J, Aji A F, Saffari A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering[J]. arXiv preprint arXiv:2306.04136, 2023.

[39] Wu Y, Hu N, Qi G, et al. Retrieve-rewrite-answer: A KG-to-text enhanced LLMS framework for knowledge graph question answering[J]. arXiv preprint arXiv:2309.11206, 2023.

[40] Wen Y, Wang Z, Sun J. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models[J]. arXiv preprint arXiv:2308.09729, 2023.

[41] Naveed H, Khan A U, Qiu S, et al. A comprehensive overview of large language models[J]. arXiv preprint arXiv:2307.06435, 2023.

[42] Wang X, Salmani M, Omidi P, et al. Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models[J]. arXiv preprint arXiv:2402.02244, 2024.

[43] Agrawal G, Pal K, Deng Y, et al. AISecKG: knowledge graph dataset for cybersecurity education[M]//AAAI-MAKE 2023: Challenges Requiring the Combination of Machine Learning 2023. 2023.

[44] 白如江,陈启明,张玉洁,等.基于 ChatGPT+Prompt 的专利技术功效实体自动生成研究[J].数据分析与知识发现, 2024,8(04):14-25. (Bai Rujiang, Chen Qiming, Zhang Yujie, et al. Research on Automatic Entities Generation of Patent Technology Function Matrix based on ChatGPT+Prompt[J]. Data Analysis and Knowledge Discovery,2024,8(04):14-25.)

[45] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[46] Dai Z, Wang X, Ni P, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]//2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei). IEEE, 2019: 1-5.

[47] Cai X, Sun E, Lei J. Research on application of named entity recognition of electronic medical records based on BERT-IDCNN-CRF model[C]//Proceedings of the 6th International Conference on Graphics and Signal Processing. 2022: 80-85.

[48] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[49] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[50] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[51] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[52] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
[53] Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert[J]. arXiv preprint arXiv:1904.09675, 2019.

通讯作者（Corresponding author）： 曹智勋（Cao Zhixun），ORCID：0009-0000-3118-5632，E-mail:zxuncao@mails.ccnu.edu.cn。

基金项目：本文系国家社科基金重大项目“数字政府建设成效测度与评价的理论、方法及应用研究”(项目编号：23&ZD081)和中央高校基本科研业务费（项目编号：CCNUJCPT2024003701）研究成果之一。
This work was supported by the Major Project of the National Social Science Fund of China "Theoretical, Methodological, and Application Research on the Measurement and Evaluation of Digital Government Construction" (Grant No. 23&ZD081) and the Fundamental Research Funds for the Central Universities (Grant No. CCNUJCPT2024003701).

作者贡献声明：
陈静：提出研究思路，提供修改意见；
曹智勋：实验设计，起草和修改论文。

利益冲突声明：
所有作者声明不存在利益冲突关系。

支撑数据：
[1]曹智勋. 大语言模型在中医知识问答领域的改进方法. DOI:10.57760/sciencedb.j00133.00319.